3А

УДК 004.89

А.В. Анисимов, К.С. Лиман, Р.И. Лупийчук, А.А. Марченко
Киевский национальный университет им. Т. Шевченко, Украина
ava@unicyb.kiev.ua, lymonadd@gmail.com, roman.lupiychuk@gmail.com,
rozenkrans@yandex.ua

Модели оценивания семантической схожести естественно-языковых текстов с использованием онтологической базы знаний WordNet

В данной статье описывается разработанная модель оценивания семантической схожести естественно-языковых текстов, которая использует в качестве базы знаний онтологию WordNet. Основной чертой данной модели является использование кластерного анализа, а также возможность одновременно обрабатывать несколько текстов. Проводится сравнительный анализ с другими схожими моделями.

Введение

Проблема вычисления оценки схожести текстов, или отнесения текстов к той или иной теме, является очень актуальной в современном мире, в силу увеличения электронного документооборота. Также данная проблема появляется при исследовании когнитивных структур лингвистического мышления.

На данный момент наиболее распространенными являются модели, основанные на сравнении словарного наполнения тестов, без подключения семантики в виде онтологических словарей и баз знаний. В данной работе как раз приводится разработанная модель оценивания схожести текстов, которая реализуется на онтологиях. Проблема словарных методов в том, что они не учитывают такие языковые феномены, как омонимия, синонимия и полисемия, когда онтологии могут помочь вовлечь эти явления в алгоритмический анализ языка.

Для семантических методов сравнения текстов, естественно, нужны методы семантического сравнения слов для определения семантической дистанции между ними. Поэтому данная статья сначала дает краткое описание той онтологии, на которую ориентируются описанные методы – это принстонский WordNet, а затем, во втором разделе статьи, приводятся описания некоторых методов определения семантической схожести слов. В последнем, третьем разделе, описаны единственная известная в литературе модель семантического сравнения текстов, а также разработанная авторами модель и небольшой теоретический сравнительный анализ.

Целью данной работы является разработка метода семантического сравнения текстов, основанного на анализе результата кластеризации семантического профиля текста, а также теоретическое сравнение данного метода с представленными в литературе.

1 Онтологическая база знаний WordNet

Дж. Миллером и его коллегами из Лаборатории когнитологии Принстонского Университета (США) была разработана модель ментального лексикона человека. Ресурс, который стал первой реализованной глобальной онтологической сетью, получил название WordNet [1] и со временем стал одним из наиболее авторитетных и распространенных стандартов, используемых для построения лексико-семантических баз.

Популярность и широкое распространение WordNet обусловлены прежде всего его существенными содержательными и структурными характеристиками. Принстонский WordNet и все последующие варианты для других языков направлены на отображение состава и структуры лексической системы языка в целом, а не отдельных тематических областей. Нынешняя версия WordNet охватывает общеупотребительную лексику современного английского языка – более 120 000 слов.

Базовой структурной единицей Принстонского WordNet является синонимический ряд (синсет), объединяющей слова с подобным значением. Каждый синсет представляет в словаре некоторое лексикализированное понятие данного языка. Для удобства использования словаря человеком каждый синсет дополнен дефиницией (gloss) и примерами употребления слов в контексте. Синсеты в WordNet связаны между собой такими семантическими отношениями, как гипонимия (родовидовое), меронимия (часть-целое), лексический вывод (каузация, пресуппозиция) и др.; среди них особую роль играет гипонимия: она позволяет организовывать синсеты в иерархические структуры (дерева, таксономии). Лексика каждой части речи представлена в виде набора деревьев (леса). Для разных частей речи родовидовые отношения могут иметь дополнительные характеристики и различаться областью распространения.

Путем между двумя синсетамы на WordNet назовем последовательность синсетов, в которой каждая последовательная пара синсетов связана определенным отношением.

2 Семантическая схожесть слов

Для получения оценки семантической схожести (или дистанции) слов было предложено много методов [2]. Те из них, которые используют онтологии, можно поделить на три группы: основанные на путях, основанные на описаниях и основанные на информационном контенте. Первые, в основном, ищут кратчайший путь в таксономии онтологии, а затем определенным образом преобразовывают полученный результат. Вторые основываются на идее, что два слова тем более похожи, чем больше у них общих слов в их словарном описании. Третьи же пытаются исправить естественный изъян таксономий – различное семантическое расстояние между понятиями, между которыми одно таксономическое звено. Например, в WordNet между понятиями FORK и SALAD FORK и между FAUNA и CHORDATE одинаковое таксономическое расстояние – одна связь типа IS-A (быть чем-либо, конкретизация, основная таксономическая связь), но интуитивно понятия из первой пары гораздо ближе друг к другу, чем со второй. Решение данной проблемы достигается благодаря введению понятия информационного контента, которое является статистической мерой специфичности того или иного слова.

Эти онтологические меры определены на множестве концептов базы знаний. Меру для слов можно получить по следующей формуле:

$$sim_X^{words}(w_1, w_2) = \max_{\substack{c_1 \in w_1.meanings \\ c_2 \in w_2.meanings}} (sim_X^{concepts}(c_1, c_2)), \quad (1)$$

где X – название меры, $w.meanings$ – это множество смыслов-концептов этого слова.

Далее приведены несколько примеров из первой и второй группы.

PATH

Простейшей, основанной на путях мерой, является мера, которую будем обозначать *PATH.* Согласно этому подходу, мерой семантической схожести между двумя концептами является обратное значение длины кратчайшего пути в таксономии между этими концептами.

$$sim_{PATH}(c_1, c_2) = \frac{1}{ShortestDist(c_1, c_2)} \ . \qquad (2)$$

LCH

Следующая мера описание которой приводится здесь, была предложена в [3]. Эту, основанную на путях меру семантического сходства, будем обозначать как *LCH.*

При этом подходе, мера сходства двух концептов определяется как отношение кратчайшего пути в IS-A иерархии к диаметру таксономии. Для WordNet 2.1 диаметр таксономии существительных равняется 17. Следующая формула описывает меру:

$$sim_{LCH}(c_1, c_2) = -\log \frac{ShorestDist(c_1, c_2)}{2 \times D} \qquad (3)$$

ShortestLength (c_1, c_2) – длина кратчайшего пути (с наименьшим количеством узлов) между концептами c_1 и c_2, а D – это диаметр таксономии (расстояние от самого общего к самому конкретному).

RES

Данная мера относится к третьей группе. Она была предложена в [4]. Для начала сформулируем понятие информационного контента, которое будем обозначать IC. Как уже говорилось выше, IC – это мера информационной специфичности концепта-синсета: чем специфичнее концепт для данного текста, тем больше его IC-значение. Таким образом, с помощью информации об употребляемости концепта, делается попытка нивелировать различие в таксономических переходах. Формально это определяется следующим образом. Расширим таксономию следующей функцией $p : C \to [0,1]$, такой, что $\forall c \in C$ $p(c)$ – вероятность встретить частичный случай концепта c в тексте. Следовательно, $p(c)$ монотонно возрастающая при движении по иерархии таксономии вверх: если $c_1\ IS-A\ c_2$, то $p(c_1) \le p(c_2)$, и $p(root) = 1$, где *root* – корень таксономии, самый общий узел.

Теперь информационный контент можно определить следующим образом:

$$IC(c) = -\log p(c) \, . \qquad (4)$$

И мера семантической похожести концептов, основанная на IC, которую будем обозначать RES, определяется в [4] так:

$$sim_{RES}(c_1, c_2) = IC(LCS(c_1, c_2)) \, , \qquad (5)$$

где LCS(c_1, c_2) – это ближайший общий родовой узел концептов c_1, c_2.

3 Семантическая схожесть текстов

Так как методы решения проблемы сравнения тематики и семантики текстов развиваются уже относительно давно, то было предложено множество подходов: Support Vector Machines, Latent Semantic Analysis, Latent Dirchlet Allocation [5], [6] и много других. Но все эти методы используют только слова, не подключая онтологические базы знаний, таким образом не учитывая такие языковые феномены, как омонимия, синонимия и полисемия. Отметим, что как в чисто «словесных», так и в «семантических» методах для представления текстов в основном используется модель «мешок со словами». То есть остаются только значимые слова (существительные), и не учитывается порядок слов, но учитывается их количество и распределение по документам. Таким

образом, синтаксическая информация отбрасывается, но статистическая и семантическая (в случае методов, которые используют онтологии) остается. «Словесные» методы часто представляют текст в виде вектора, где каждый элемент вектора представляет слово в этом документе. Самой распространенной схемой можно назвать TF-IDF (Term Frequency – Invert Document Frequency):

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \tag{6}$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}, \tag{7}$$

$$tfidf_{ij} = tf_{ij} \times idf_i, \tag{8}$$

где i – индекс терма (слова), j – индекс документа в коллекции, tf_{ij} – частота терма i в документе j, idf_i – количество документов, в которые входит терм i, $tfidf_{ij}$ – вес терма i в документе j.

После такого представления легко применять многие математические методы, например, кластеризацию с помощью SVM в задаче разделения документов по темам. В этом случае SVM пытается строить разделяющую плоскость между двумя множествами точек.

Методы оценивания семантической схожести, рассмотренные в этой статье, определены на онтологиях, которые изначально представляют граф. Строго говоря, графоподобная структура онтологии не является необходимой, достаточно, чтобы на базе знаний, которая используется, была определена мера схожести (а точнее, семантическое расстояние) между концептами. Далее будем предполагать, что используется мера PATH (2).

Семантическим профилем текста будем называть взвешенный подграф онтологии, который является проекцией слов текста на граф онтологии: каждому слову ставятся в соответствие его вершины-синсеты из онтологии. Вес слов подграфа в каждой модели рассчитывается по TFIDF (8), а вес вершин – по-разному.

3.1. Метод потока через сеть

Метод потока через сеть, или NFM (Network flow method), был описан в [5]. Идея этого метода состоит в том, чтобы найти значение минимального потока между двумя семантическими профилями (каждый представляет определенный текст). «Цена» этого потока и будет значением схожести текстов.

Более формально. Пусть $G =| (N, E)$ – некоторый граф (онтология). N – вершины (концепты), E – ребра (отношения между концептами). Каждое ребро имеет вес $c : E \to \mathrm{R}$, который равен онтологической дистанции между концептами (раздел 2). Схема расстановки весов в этой модели следующая. Каждая вершина $i \in N$ имеет вес $b(i) : N \to \mathrm{R}$, который определяет, является вершина источником ($b(i) > 0$), или стоком ($b(i) < 0$), или указывает, что вершина не задействована ($b(i) = 0$). Следовательно, задача состоит в том, чтобы найти минимальный поток от источников к стокам. Один текст обозначим источником (устанавливаем положительные веса для вершин семантического профиля), а второй – стоком (устанавливаем отрицательные веса). Для общих концептов считается сумма соответствующих весов. То есть возможна ситуация, когда вес одной вершины уравновесится и станет равным 0, то есть понятие, которому соответствует вершина, одинаково представлено в обоих текстах, а значит, и не выражает их особенностей. При этом веса вершин нормируются таким образом, чтобы общий исток был равен общему стоку.

Далее нужно решить задачу поиска наименьшего потока. Пусть IN_i будет множеством ребер *(h,i)*, через которые поток входит в вершину *i*; аналогично OUT_i будет множеством ребер *(i,j)*, через которые поток выходит из *i*. Тогда поток, проходящий через вершину *i*, описывается функцией $x: E \to \mathrm{R}$. Допустимым решением будет поток *x*, такой, что разница входного ($\sum_{(h,i)\in IN_i} x(h,i)$) и выходного ($\sum_{(i,j)\in OUT_i} x(i,j)$) потоков будет равна предложению или потреблению вершины (*b(i)*). Формально проблема поиска наименьшего потока может быть выражена следующим образом:

$$\sum_{(i,j)\in E} c(i,j)\cdot x(i,j) = z(\vec{x}) \to min \qquad (9)$$

со следующими ограничениями:

$$\sum_{(i,j)\in OUT_i} x(i,j) - \sum_{(h,i)\in IN_i} x(h,i) = b(i),\ \forall i \in N, \qquad (10)$$

$$x(i,j) \ge 0,\ \forall (i,j) \in E. \qquad (11)$$

3.2 Модель, основанная на кластеризации

Разработанная модель (далее МОК) определения семантической похожести текстов основана на том факте, что если текст посвящен какой-то определенной теме (или имеет четко выраженные темы), то лексика этого текста, а следовательно, и семантический профиль, будут элементы с сильно выделяющимся весом. Мало того, используя возможность измерить семантическое расстояние на онтологии, можно выделить группы синсетов, которые будут плотно сгруппированы. Эти группы (если мы спроектируем синсеты на гиперплоскость, с сохранением расстояний, то можно говорить об областях) близких синсетов указывают на скопление синонимов или понятий относящихся к одной теме.

Теперь рассмотрим ситуацию с двумя текстами T_1 и T_2 и соответственно с двумя семантическими профилями SP_1 SP_2. В этом случае схема взвешивания вершин-синсетов будет следующая. Сначала веса слов в текстах нормализируются по размеру текстов. Вес синсетов вычисляется как сумма соответствующих ему слов из обоих текстов, но при этом запоминается вклад каждого текста.

$$SP_1 = \{s_{11}, s_{12}, \ldots, s_{1n_1}\};\ SP_2 = \{s_{21}, s_{22}, \ldots, s_{2n_2}\}, \qquad (12)$$

где s_i – следующая структура

$$s_i = \left\langle \sum_{T_1,T_2} w(i) \middle| Contribution = \{T_1(i), T_2(i)\} \right\rangle, \qquad (13)$$

Теперь, после кластеризации множества синстетов на плотные и тяжелые области, которые олицетворяют темы, можно проанализировать вклад каждого текста в данную тему или наоборот – вычислить присутствие данной темы в том или ином тексте. Тему в данном случае удобно представить как нечеткое множество:

$$Th_j = \{(s_i, \tau_j(i)) \,|\, s_i \in \bigcup_k SP_k\}, \qquad (14)$$

где $\tau_j(i)$ – это функция принадлежности *i*-го синсета к *j*-й теме.

3.3 Анализ и тестирование

Представленная в этой статье модель определения семантической схожести текстов теоретически может быть применима сразу к нескольким текстам, в отличие от многих других моделей. Естественно, при слишком большом количестве текстов или при попытке анализа очень больших, всеохватывающих текстов, метод, скорее всего, покажет плохие результаты, как, впрочем, и остальные модели в схожей ситуации.

Более точный анализ плюсов и минусов данной модели требует дополнительного исследования, так как результаты МОК сильно зависят от использованного алгоритма кластеризации. То есть необходим определенный подход, который будет учитывать именно семантико-лингвистическую силу той или иной модели, с целью выявить ту модель, которая будет более других соответствовать когнитивным структурам. Другой путь сравнения вышеописанных и прочих методов оценивания схожести текстов – это проверка в различных прикладных и тестовых задачах, как например, была протестирована модель NFM в [5], где показала существенное улучшение показателей в некоторых задачах, что вселяет надежду на семантический подход к оценке схожести текстов. Но при этом также важно вычленить вклад собственно модели схожести текстов и решателя конкретной задачи. На данный момент это является темой последующих исследований.

Выводы

В данной статье была представлена модель определения семантической схожести естественно-языковых текстов, основанная на кластеризации. Данная модель имеет определенные преимущества относительно других моделей, но более точное сравнение требует более глубокой разработки теоретической базы.

Литература

1. Режим доступа : http://wordnet.princeton.edu/
2. Анисимов А.В. Методы вычисления мер семантической близости слов естественного языка / А.В. Анисимов, К.С. Лиман, А.А. Марченко // Искусственный интеллект. – 2010. – № 3.– С.170-176.
3. Leacock C. Combining local context and WordNet similarity for word sense identification / C. Leacock, Chodorow // Fellbaum, C. WordNet: An electronic lexicaldatabase. MIT Press. – M., 1998. – P. 265-283.
4. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy / Resnik P. // Proceedings of the 14th International Joint Conference on Artificial Intelligence. – 1995. – P. 448-453.
5. Tsang V. A Graph-Theoretic Framework for Semantic Distance / V. Tsang, S. Stevenson // Computational Linguistics. – 2010. – Vol. 36. – P. 31-69.
6. Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, 1998.

Literatura

1. http://wordnet.princeton.edu/
2. Anisimov A.V. Iskusstvennyj intellekt. Vyp. 3. 2010. S. 170-176.
3. Leacock CFellbaum C. ed.WordNet: An electronic lexical data base. MIT Press. 1998. P. 265-283.
4. Resnik P. Proceedings of the 14th International Joint Conference on Artificial Intelligence.1995. P. 448-453.
5. Tsang V. Computational Linguistics. Vol. 36. 2010. P. 31-69.
6. JoachimsT. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. 1998.

А.В. Анісімов, К.С. Лиман, О.О. Марченко Р.І. Лупійук

Моделі оцінювання схожості природно-мовних текстів з використанням онтологічної бази знань WordNet

В даній статті описується розроблена модель оцінювання семантичної схожості природно-мовних текстів, яка використовує в якості бази знань онтологію WordNet. Основною рисою даної моделі є використання кластерного аналізу, а також можливість одночасно обробляти декілька текстів. Наводиться порівняльний аналіз з іншими моделями такого ж ґатунку.

A.V. Anisimov, K.S. Lyman, A.A. Marchenko, R.I. Lupijchuk

Estimation Models of Semantic Similarity of Natural Language Texts with Using Ontological Knowledge Base WordNet

The developed model of text semantic similarity estimation that uses WordNet ontology as knowledge base is described in this article. The main feature of this model is involving of the cluster analysis and multitext processing ability. The comparative analysis of this model to other similar models is performed.

Статья поступила в редакцию 30.06.2011.